ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka

Shilpa Ankalaki[1], Neeti Chandra[2], Jharna Majumdar [3]

Asst. Prof, Department of M.Tech CSE, Nitte Meenakshi Institute of Technology, Bangalore, India[1]

Student, Dept. of M Tech CSE, Nitte Meenakshi Institute of Technology, Bangalore, India, [2]

Dean R&D, Prof. & Head Dept. of M. Tech CSE, Nitte Meenakshi Institute of Technology, Bangalore, India[3]

Abstract: We have entered the era of Big Data where data is emerging with 5V (Velocity, Variety, Volume, Value, Veracity), making it complex and useful for the predictive and descriptive analysis. Decision making with analysis is an important concern in modern agriculture. One such decision making process is related to the forecast of crop yield in various environmental and soil conditions. This basis of the work is based on this data mining process. The work deals with two subtasks, first, it implements and compares different clustering method for the districts having similar kind of productivity factors for crops, and second, forecasting the yield of the major crops for different districts.

Keywords: Batchelor & Wilkins', DBSCAN, AGNES, Multiple Linear Regression.

I. INTRODUCTION

For any country national economy, agriculture plays a significant role. Agriculture is a task full of risks which is influenced by several factors like temperature, rainfall, sown area, past production record and other soil and climate related issues. Reliable information about these factors can be helpful to farmers as well as government in decision making.

a. It helps farmers in providing the historical crop yield record with a forecast reducing the risk management.
b. It helps government in making crop insurance policies as well as policies for supply chain operation.

Apart from these, data mining can be found as an effective tool in the applications like weather forecasting, drought management, yield prediction etc. Crop yield forecasting can be done using various data mining techniques. In proposed work, data mining techniques like Batchelor Wilkin's, DBSCAN, AGNES, and regression methods are used to forecast the annual yield of major crops.

II. LITERATURE SURVEY

The objective is to enhance the DBSCAN in order to detect the cluster on its own by finding the input parameter. [1] The difficulty comes in discovery of clusters, as it did not execute well on multi-density data sets. For determining the "Epsilon (Eps)" value, first the k-distance graph for all points is needed to be drawn, which will be entered by the user. The average distance to the all the k points will be calculated. The aim to determine the "knee" for the estimation of the Eps parameter. For Min-pts: The total of data entities in "Eps- neighborhood" of all point in data-set is calculated in succession. The paper deals with an analysis of the prediction of rice production in Bangladesh. [2] Bangladesh offers several varieties of rice which have different cropping season. For this a prior study of climate (effect on temperature and rainfall) in Bangladesh and its effect on agricultural production of rice has been done.

Then this study was being taken into regression analysis with temperature and Rainfall. Temperature puts an adverse consequence on the crop production. The data has been taken from the "Bangladesh Agricultural Research Council (BARC)" for past 20 years with 7 attributes: "rainfall", "max and min temperature", "sunlight", "speed of wind", "humidity" and "cloud-coverage". In Pre-processing, the whole dataset was divided in 3 month duration phases (March to June, July to October, November to February).

For these duration, the average for every attribute has been taken and associated with it. In Clustering, the different pre-processed table has been analyzed to find the sharable group of region based on similar weather attribute. Based on this result classification can be done by giving it as input. In clustering, "Self-Organizing maps (SOM)", was employed. It has low classification error as well as reduces the dimension of input dataset. Then in classification from 4 regions, 2 were discarded. Finally the errors in training, testing &cross justification were dignified in terms of "Root mean-square error (RMSE)".

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

Data mining is a strong tool for analysis of large datasets which is primarily distributed in two groups i.e. "Classification" and "Clustering". [3] Clustering is based on similarity of data, while classification is for classifying unknown samples. Density based clustering divides the data in non-equal cluster based on Euclidean distance.

In this MLR technique and density based technique has been compared for the estimation of production. The parameters considered are Year, area of sowing, rainfall, yield/acre, exact production.

III. DATA MINING APPROACH IN AGRICULTURE

An agricultural data set consist of several factors like area, production, temperature, rainfall etc. having values very close to each other. For proposed work, agricultural dataset has been taken from several sources some of the links are as follows: https://data.gov.in/, http://raitamitra.kar.nic.in/statistics.html#CROPWISE_NORMAL_AREA,http://14.139.94.101/fertimeter/Distkar.aspx, http://raitamitra.kar.nic.in/ENG/statistics.asp, http://dmc.kar.nic.in/trg.pdf. Different clustering can be formed from which descriptive analysis as well as association rules can be made from the clustering. Different clustering algorithms are available but which clustering is good for these kind of datasets where the data values related to the instances in data set are very close. A comparative study is required for this.

A. Modified DBSCAN with Batchelor& Wilkins' Determination of Eps and Minpts

This algorithm is built on the idea that for a data point to form a cluster there must be a pre-defined least number of neighboring data points in a given radius. This takes two input parameter.

I. Epsilon (Eps), the radius of neighborhood
II. Minimum points (Minpts): Minimum count of points that is to be considered to model a cluster.

The Epsilon (Eps) value can be found by drawing a "K-distance graph" for entire data-points in dataset for a given 'K', entered by the user. But the 'K' value is a user input. To overcome this user dependency Batchelor Wilkins' algorithm is used. The "Batchelor& Wilkins' algorithm will give the number of clusters which can be used in "K-distance graph". [1]

At first, the distance of a point to every 'K' of its nearest-neighbors is calculated. The calculated value is averaged and sorted and the graph is plotted. When the graph is plotted, a knee point is determined in order to find the optimal Eps value. The Minpts is calculated by dividing the number of points within the Eps - neighborhood by the total count of data points. Let the total count of data-points in dataset be D and the total points inside Eps neighborhood is N.

ALGORITHM:
DBSCAN [4]
INPUT: Dataset (X1….Xn), Eps, Min-pts
OUTPUT: Clusters of data points(X1….Xn)

STEPS:
1. Select an arbitrary point, say X1 which is not visited.
2. Find the distance of remaining data points to the selected point. Select the points whose distance is less than Eps.
3. Count the number of data point, say N, having distance less than Eps. If the N is less than or equal to Min-pts, mark the N points as visited, else, as noise.
4. Choose the point found as a part of cluster in step iii and repeat the process from step ii until all points in the cluster is determined.
5. If (any new unvisited node) Repeat the process from 2 to 4. Else the process is terminated

B. AGNES

ALGORITHM: Agglomerative Nesting (AGNES) [5]
This clustering is based on the idea that nearby data points results in same clusters
INPUT: Dataset (X1,X2……..Xn), distance metrics, linkage method
OUTPUT: Cluster of data object.

STEPS:
1. Compute the distance matrix for the object features of given dataset D.
2. Take all the data points as individual cluster i.e. "n" clusters.
3. Repeat Until only one cluster is left

a. Find the distance metrics for the clusters and find the pair of clusters which is closest cluster min D(Ci, Cj).
b. Merge Ci and Cj into a one new cluster say Ci,j. Remove the Ci and Cj from the set of n clusters and add Ci,j to the set.

IV. REGRESSION ANALYSIS TO FORECAST CROP YIELD

The crop yield forecasting can be done by estimating the production for the upcoming year based on the past year production values and other production dependent variables. A linear expression is derived from the regression model based on which the production is estimated. Based on the estimated value farmers can decide whether care is needed for the crop in the initial stage or not. Even the government can make policies based on that. If production is estimated less, crop insurance policy can be made. If production is estimated more, policies regarding market price as well as decision on exporting can be made. When implementing regression there are several independent variables. But to determine the significance of variables in determining the dependent value is useful as to

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

get rid of unnecessary calculations. The p-value tests the null hypothesis i.e. the term will have no effect on the regression equation. An assumption is made that the confidence level of the regression will be 95%.
An independent variable which has a "p- value" of less than 0.05, specifies that the "null-hypothesis" can be rejected means it will have effect on regression analysis. So these independent values can be added to the model as changes in the value of these variables will result in change of dependent variable. Whereas if the p-value is more than common alpha level i.e. 0.05, the variable will said to be not significant to the model.

For a given dataset where $X_1 \ldots X_k$ are independent variables and Y is a dependent variable, the multiple linear regression fits the dataset to the model:

$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \ldots + \beta_k x_{k_i} + \varepsilon_i$$

Where β_0 is the y-intercept and the parameters β_1, β_2 , β_3 ..., β_κ are called the partial coefficients.

V. EXPERIMENTAL RESULTS

A. DBSCAN
This algorithm is provided with the input in two ways: one for large dataset having 2446 instances with 13 attributes like area, production, temperature, rainfall, pH, Soil minerals(Nitrogen, phosphorus, potassium) etc. The main idea behind this is to get a comparative analysis of clustering algorithm suitable for agricultural data and the efficiency of the implemented methods follow different similarity criterion, the results also differ accordingly. To find out the number of clusters automatically Batchelor Wilkin's algorithm has applied on the agriculture database of instances 2446. As a result 7 clusters are formed. It will be input to the KNN plot (K=7) to find the optimal minimum number of points. Fig1 shows result of the KNN (7NN for K=7). It shows that 0.4 as optimal minimum radius for DBSCAN.

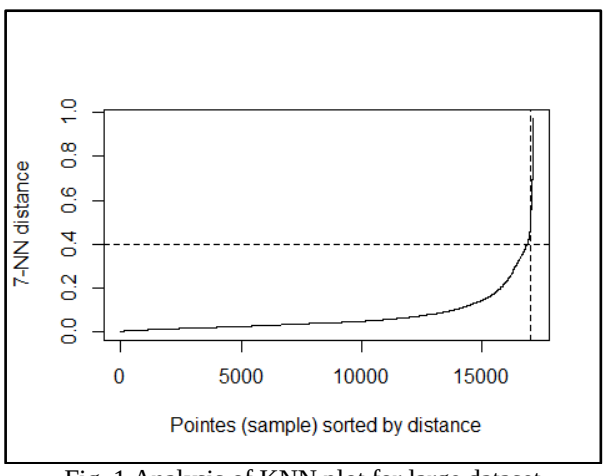

Fig. 1 Analysis of KNN plot for large dataset

TABLE I COMPARISON DBSCAN CLUSTERING FOR AGRICULTURE DATASET

Analysis of DBSCAN clustering for Agriculture Dataset		
Epsilon	0.4	0.3
Min-points	7	1
Noise	1	1
Cluster	2	7
Execution Time(in Sec)	35.1373607	16.089070

B. Agglomerative Nesting (AGNES)
As AGNES algorithm is a hierarchical algorithm, to form a hierarchy the full dataset cannot be passed because the dendogram it will create will be too cluttered that any inference can't be drawn from it. So for better inference and conclusion, the averaged data set and related attribute have been taken. As the different methods follow different similarity criterion, the results also differ accordingly.

From the comparison of cluster distribution plot shown in Fig. 2, it can be inferred that Average, Complete and Wards method shows a normally distributed clusters. While the execution time comparison plot shown in Fig. 3, is showing that single linkage is having the best execution time highlighted with green bar and wards method having worst execution time highlighted by red bar.

TABLE III COMPARISON OF LINKAGE METHODS (NUMBER OF DATA POINTS AND EXECUTION TIME)

	Cluster 1	**Cluster 2**	**Cluster 3**	**Execution Time**
Average	19	5	3	5.86934
Single	24	2	1	3.85522
Complete	15	10	2	6.31136
Centroid	25	1	1	6.63938
Ward	19	5	3	6.67638

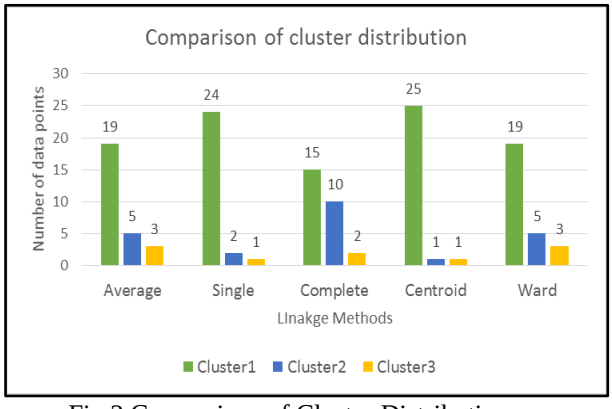

Fig.2 Comparison of Cluster Distribution

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

As a conclusion drawn it can be said that average linkage, gives the best result. From Fig. 4, complete linkage can be seen as giving the best results.

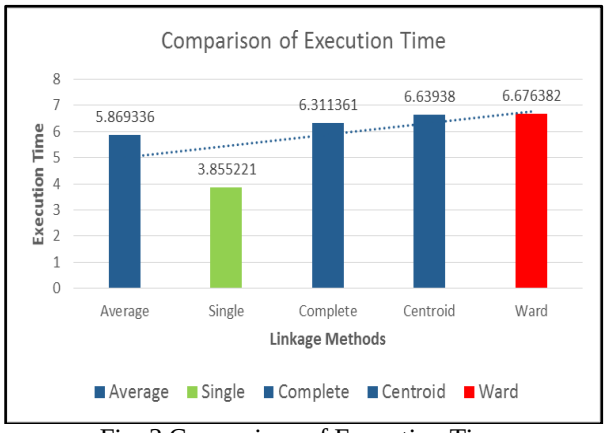

Fig. 3 Comparison of Execution Time

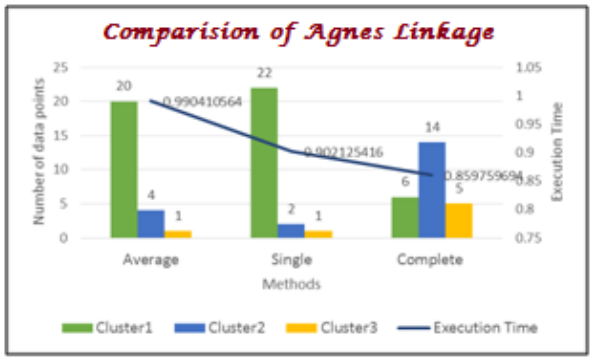

Fig. 4 Comparison of different linkage methods of Agnes Clustering

C. Multiple Linear Regression for forecasting of crop. Significance Test

The highlighted cells are representing the insignificant independent attributes for each crop as the values are more than 0.05. Regression Equation is formed using the independent variables.

TABLE IIIII SIGNIFICANCE TEST FOR INDEPENDENT VARIABLES

	Cotton	Groundnut	Jowar	Rice	Wheat
Temperature	**0.547536**	3.41E-07	3.86E-07	0.003139	0.001137
Rainfall	**0.784625**	1.86E-06	**0.653187**	**0.105878**	0.018042
pH	0.011752	2.55E-05	0.029733	5.08E-07	0.01834
Nitrogen	5.85E-05	**0.071873**	**0.349257**	0.000841	8.6E-06
Phosphorus	**0.071843**	0.043345	**0.464847**	0.025816	**0.209524**
Potassium	2.82E-07	**0.643528**	0.050831	1.43E-05	0.021422
Water	4.95E-05	4.92E-49	1.2E-102	1.22E-26	NA

TABLE IVV MULTI LINEAR REGRESSION EQUATION FOR CROP WISE YIELD

Crop	Yield Forecast Equation
Cotton	Yield=(7.149372)+(-0.14468)pH+ (-0.00131)Nitrogen+ (-0.00405)Potassium+(-0.00405) Water Required
Ground nut	Yield=(2.79115)+(0.029217)Temperature+(5.78e-05)Rainfall+(-0.05681)pH+(0.00127) Phosphorus+(-0.00492)Water Required
Jowar	Yield=(-1.62694)+ (-5.35e-02)Temperature+(0.051512)pH+ (-0.00113) Potassium+(0.01685436)Water Required
Rice	Yield=(-0.18503)+(0.041593)Temperature+(0.172042)pH+(-8.27e-04)Nitrogen+(-4.28e03) Phosphorus+ (-0.00264)Potassium+(9.15e-04)Water Required
Wheat	Yield=(112)+(-4.14e-02)Temperature+ (1.34e-04)Rainfall+(0.079153)pH+(-1.31e-03) Nitrogen+ (-0.00167)Potassium+(-0.28125)Water Required

VI. CONCLUSION

From the different comparisons performed between the clustering algorithms it is very clear that the effectiveness of clustering algorithm is data dependent. It means clustering algorithm A can be good with dataset X while B can be more efficient with dataset Y. From the proposed work we can conclude that DBSCAN is more time consuming than the Agnes and Agnes with Average links gives the optimal and efficient number of clusters.

Regression analysis performed for the forecasting shows a highly dependency on the dataset. If the data collected is significant then the results will fit the model. Otherwise it can lead to some imprecise results.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

VII. FUTURE WORK

Use of such kind of approach to forecasting is not restricted to agriculture alone. The clustering and regression is one of the capable tool in field of data-mining which can be used in several different ways. The clustering can also be implement in the concept of soil type clustering so that soils having similar kind of features can be used for similar kind of crops. The concept can be further merged with the market data to predict the price of crop, as well as to predict the fertilizers consumption. This is not limited to agriculture; the concept can be deployed in weather forecasting also. All these combined together can be a very good work in the field of precision agriculture.

ACKNOWLEDGMENT

The authors express their sincere gratitude to **Prof. N. R Shetty,** Advisor and **Dr. H C Nagaraj,** Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT. The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

REFERENCES

[1] Manisha Naik Gaonkar, KedarSawant, "AutoEpsDBSCAN: DBSCAN with Eps Automatic for Large Dataset", IJACTE, 2013.
[2] Mohammad Motiur Rahman, Naheena Haq, Rashedur M Rahman, "Application of data mining tools for rice yield prediction on clustered regions of Bangladesh", IEEE, 2014, Page(s):8 – 13.
[3] B Vishnu Vardhan, D. Ramesh, O Subhash Chander Goud, "Density Based Clustering Technique on Crop Yield Prediction", International Journal of Electronics and Electrical Engineering Vol. 2, No. 1, March, 2014.
[4] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
[5] K.Sasirekha, P.Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013.
[6] http://www.kdnuggets.com/2014/08/four-main languages-analytics-data-mining-data science.html
[7] Zhijie Xu, Laisheng Wang, Jiancheng Luo, Jianqin Zhang, "A modified clustering algorithm for data mining", IEEE, 2005.
[8] Raorane A.A, Kulkarni R.V, "Data Mining: An effective tool for yield estimation in the agricultural sector", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2012.
[9] M.C.S.Geetha, "A Survey on Data Mining Techniques in Agriculture", International Journal of Innovative Research in Computer and Communication Engineering, 2015.
[10] Anshal Savla, Nivedita Israni, Parul Dhawan, Alisha Mandholia, Himtanaya Bhadada, Sanya Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture", IEEE, 2015, Page(s):1 – 7.
[11] K.Sasirekha, P.Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review", International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013.
[12] Dr. A.Bharathi, E.Deepankumar, "Survey on Classification Techniques in Data Mining, IJARCCE, 2009.